Breast Cancer Research

RESEARCH

Open Access

A triple hormone receptor ER, AR, and VDR signature is a robust prognosis predictor in breast cancer

Mohamed Omar[1,2], J. Chuck Harrell[3], Rulla Tamimi[1,4], Luigi Marchionni[1], Cihat Erdogan[5], Harikrishna Nakshatri[6] and Tan A. Ince[1,7*]

Abstract

Background Despite evidence indicating the dominance of cell-of-origin signatures in molecular tumor patterns, translating these genome-wide patterns into actionable insights has been challenging. This study introduces breast cancer cell-of-origin signatures that offer significant prognostic value across all breast cancer subtypes and various clinical cohorts, compared to previously developed genomic signatures.

Methods We previously reported that triple hormone receptor (THR) co-expression patterns of androgen (AR), estrogen (ER), and vitamin D (VDR) receptors are maintained at the protein level in human breast cancers. Here, we developed corresponding mRNA signatures (THR-50 and THR-70) based on these patterns to categorize breast tumors by their THR expression levels. The THR mRNA signatures were evaluated across 56 breast cancer datasets (5040 patients) using Kaplan–Meier survival analysis, Cox proportional hazard regression, and unsupervised clustering.

Results The THR signatures effectively predict both overall and progression-free survival across all evaluated datasets, independent of subtype, grade, or treatment status, suggesting improvement over existing prognostic signatures. Furthermore, they delineate three distinct ER-positive breast cancer subtypes with significant survival in differences—expanding on the conventional two subtypes. Additionally, coupling THR-70 with an immune signature identifies a predominantly ER-negative breast cancer subgroup with a highly favorable prognosis, comparable to ER-positive cases, as well as an ER-negative subgroup with notably poor outcome, characterized by a 15-fold shorter survival.

Conclusions The THR cell-of-origin signature introduces a novel dimension to breast cancer biology, potentially serving as a robust foundation for integrating additional prognostic biomarkers. These signatures offer utility as a prognostic index for stratifying existing breast cancer subtypes and for de novo classification of breast cancer cases. Moreover, THR signatures may also hold promise in predicting hormone treatment responses targeting AR and/or VDR.

Keywords Breast cancer, Cell-of-origin, Triple hormone receptor, Estrogen receptor, Androgen receptor, Vitamin D receptor, Survival, Predictive modeling, Breast cancer subtypes

*Correspondence:
Tan A. Ince
tai9015@med.cornell.edu
Full list of author information is available at the end of the article

Background

The term 'classification' is often used to denote subgroups of breast cancer (BrCa) with different outcomes or treatments. However, utilizing predictive or prognostic clusters for classification has inherent instability, wherein group definition may change with treatment or survival modality. In contrast, normal cell types have been used as a stable reference point to classify hematological malignancies (Figure S1) [1, 2], which inspired us to develop a similar approach for BrCa.

We previously identified four differentiation types in human normal breast luminal epithelial (NBLE) cells based on receptor protein expression for estrogen, androgen, and vitamin-D: ER, AR, and VDR [3, 4]. Triple hormone receptor-positive (THR-3) NBLE cells co-express all three receptor proteins, while THR-2, THR-1, and THR-zero (THR-0) cells co-express two, one, or none, respectively [3, 4]. We found that each THR differentiation stage has a unique DNA methylation profile [5] and human BrCa preserve the THR DNA methylation profile of NBLE cells [6, 7]. This suggests that either human breast tumors keep the initial THR state of their singular normal cell-of-origin or that a differentiation block limits them to one dominant THR state, like hematological malignancies [3, 4].

While ER has been a key prognostic and predictive marker in BrCa [8], the role of VDR and AR in BrCa prognosis has been less clear. In multivariate analysis, VDR protein expression does not correlate with BrCa overall survival (OS) [9–12]. Similarly, the prognostic role of AR in BrCa is not well defined [13–15], with some studies linking AR expression to a better prognosis in ER-positive BrCa, and a worse prognosis in ER-negative BrCa [16, 17]. Therefore, our initial motivation for combining AR and VDR with ER was solely based on this triple marker panel's ability to identify distinct normal breast cell types, rather than their potential additive prognostic power. Nonetheless, we discovered that combining these three markers can unexpectedly form a powerful prognostic panel [3, 4, 18].

Since the 1990s, BrCa has been categorized into three clinicopathologic subgroups based on the expression of ER, PR, and the human epidermal growth factor receptor 2 (HER2). The estrogen receptor (ER) is expressed in approximately 70% of BrCa [8], and its activation by estrogen drives BrCa growth, making ER-positive tumors susceptible to anti-estrogen therapies [19, 20]. HER2 is amplified in 10–15% of BrCa that are treated with anti-HER2 therapies. The remainder of BrCa subtypes are negative for ER, PR, and HER2, which are referred to as triple-negative breast cancers (TNBC).

Several mRNA-based predictive and prognostic signatures, including MammaPrint [21, 22], Oncotype DX [23], and Prediction Analysis of Microarray 50 (PAM-50) [24] have been developed to move beyond the ER/PR/HER2 paradigm. However, these signatures have typically demonstrated clinical utility primarily in node-negative, ER-positive, and HER2-negative tumors [25, 26], thus benefitting fewer than half of the patient population [25].

In this study, we present a cell-of-origin-based analysis of BrCa. While it remains in its early stages compared to the hematopoietic system, this approach has already demonstrated significant promise for defining BrCa prognostic subtypes, as we will elaborate upon next.

Methods

Samples and inclusion criteria

Cell lines

To develop the triple hormone receptor (THR) mRNA signature, we used gene expression profiles of BrCa cell lines from the Cancer Cell Line Encyclopedia (CCLE) dataset [27], including those positive for a single hormone receptor or none with THR-[0/1] phenotype (BT-20, HCC1187, HCC1937, HCC1143, and MDA-MB-231) together with cell lines positive for two or three receptors with a THR-[2/3] phenotype (MCF7, T47D, CAMA-1, YMB-1, and ZR-75–1).

Patient data

The mRNA expression data from multiple cohorts comprising samples from BrCa patients with available survival information were analyzed. These cohorts included the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort [28] (n = 1904), KM plotter (KMP) cohort (n = 2,032, 50 studies) [29], Meta-10 cohort comprising ten different studies [30], nine of which overlap with the KMP cohort except for GSE4922 (n = 249) [31], and the BC855 cohort (n = 855, 4 studies) [21, 32–34].

Development of the triple-hormone receptor gene signatures

To identify genes reflecting THR receptor status, we conducted a t-test based differential expression analysis between THR-0/1 and THR-2/3 BrCa cell lines in the CCLE dataset [27]. The top 50 differentially expressed genes (DEGs) sorted by *p*-value were selected for further analysis, termed as THR-50 (Additional File 2).

Subsequently, we extended this analysis to human data using gene expression profiles from 855 BrCa samples in the BC855 cohort [21, 32–34]. Here, samples were categorized into THR-0, THR-1, THR-2, and THR-3 based on *ESR1*, *AR*, and *VDR* expression levels. Similar to CCLE, we performed differential expression analysis to identify genes distinguishing THR-0/1 from THR-2/3. We prioritized 350 genes common between CCLE and

BC855 cohorts, ranked by SAM fold change and q-value (in BC855) or *p*-value (in CCLE), resulting in a set of 70 genes defining THR status in both cell lines and human tissue, referred to as THR-70 (Additional File 3).

Additional details on the derivation of THR-50 and THR-70 can be found in the Supplementary Methods (Additional File 1).

Unsupervised clustering of single nucleus transcriptome data of healthy breast tissues

Single nucleus chromatin accessibility and transcriptome data of normal breast tissues of women of diverse genetic ancestry have been described recently [35]. Epithelial cells in the normal breast have recently been renamed as Basal-Myoepithelial (BM, replacing the previous name of basal cells), luminal adaptive secretory precursor (LASP, replacing the previous name of luminal progenitor cells), and luminal hormone sensing (LHS, replacing the previous name of mature luminal cells) [35]. BM cells are divided into BM-basal alpha (BM-BAα) and BM-BAβ, LASP cells into alveolar progenitor (LASP-AP) and basal-luminal hybrid (LASP-BL), and LHS cells into LHSα and LHSβ [36].

The single nucleus RNA-seq (snRNAseq) data was normalized and scaled with the *NormalizeData* and *ScaleData* functions in Seurat [37], using the default parameters. Subsequently, a heatmap of the single-nucleus gene expression data for the THR-70 genes was generated with the *DoHeatmap* function in the Seurat package, for the epithelial subtypes.

Enrichment in normal human breast epithelial cells

To further evaluate the enrichment of THR in normal breast epithelial cells, we used single cell RNA-seq (scRNA-seq) profiles of healthy breast tissue from two independent datasets [35, 38]. Specifically, we measured the expression and enrichment of THR genes across the different epithelial cell clusters using the normalized and z-scored expression profiles. Enrichment was computed using *UCell*, a gene signature scoring method based on the Mann–Whitney U statistic [39] and the computed scores were smoothed using the weighted average of the k-nearest neighbors in principal components analysis.

Survival analysis

Survival analysis was utilized to examine the relationship between THR signatures and the outcome of various patient cohorts using the Kaplan–Meier (KM) survival estimate with log-rank tests [40, 41]. Patients were stratified into two or four groups (quartiles) based on either average expression of the signature genes or risk-score calculated by logistic regression models. Additional information on the categorization process and thresholding methods can be found in the Supplementary Methods (Additional File 1). The relative hazard ratios were determined using Cox proportional hazard regression analysis, with *p* values computed using the Wald test [30].

Gene set enrichment analysis

The immune module of the Gene Set Cancer Analysis (GSCA) online platform [42] was used to determine the correlation between immune cell infiltrates and GSVA enrichment score using ImmuCellAI (Immune Cell Abundance Identifier), which estimates the abundance of 24 immune cell types [43]. Further details on the gene set enrichment analysis can be found in the Supplementary Methods (Additional File 1).

Unsupervised clustering of patient samples

To assess the capacity of THR signatures for defining robust BrCa subtypes, we conducted unsupervised hierarchical clustering of the samples within the METABRIC cohort. The clustering was performed using the ward minimum variance method [44], followed by demarcating the hierarchical tree into five groups. The optimal number of groups was determined based on the observed patterns and the clinical interpretability. Next, we evaluated overall survival (OS) and recurrence-free survival (RFS) probabilities among the five groups using KM survival curves and the log-rank test.

Software and statistical analysis

All statistical analyses were performed using R software (version 4.0.3). The *survival* and *survminer* R packages were used for generating KM survival curves and the COX proportional hazards models [45, 46]. The *stats* package was used for hierarchical clustering, while the *glmnet* package was used to fit the logistic regression models [47]. The significance level (*p*-value and FDR) was set at 0.05 for all statistical tests except for the CCLE differential expression analysis, for which we used a *p*-value threshold of 0.01.

Results

Classification of breast tumors by triple hormone receptor protein expression

The triple-hormone receptor (THR) categorization is based on protein expression of ER, AR, and VDR, assessed by immunohistochemical (IHC) staining of formalin-fixed paraffin-embedded (FFPE) BrCa tissue microarrays (TMA) as described before [3, 4], producing four subgroups: THR-0, THR-1, THR-2, and THR-3, representing 7%, 11%, 28%, and 54% of BrCa, respectively (Fig. 1A). We previously showed in KM survival analysis that BrCa with fewer hormone receptors is associated with shorter OS [3] with a statistically significant hazard

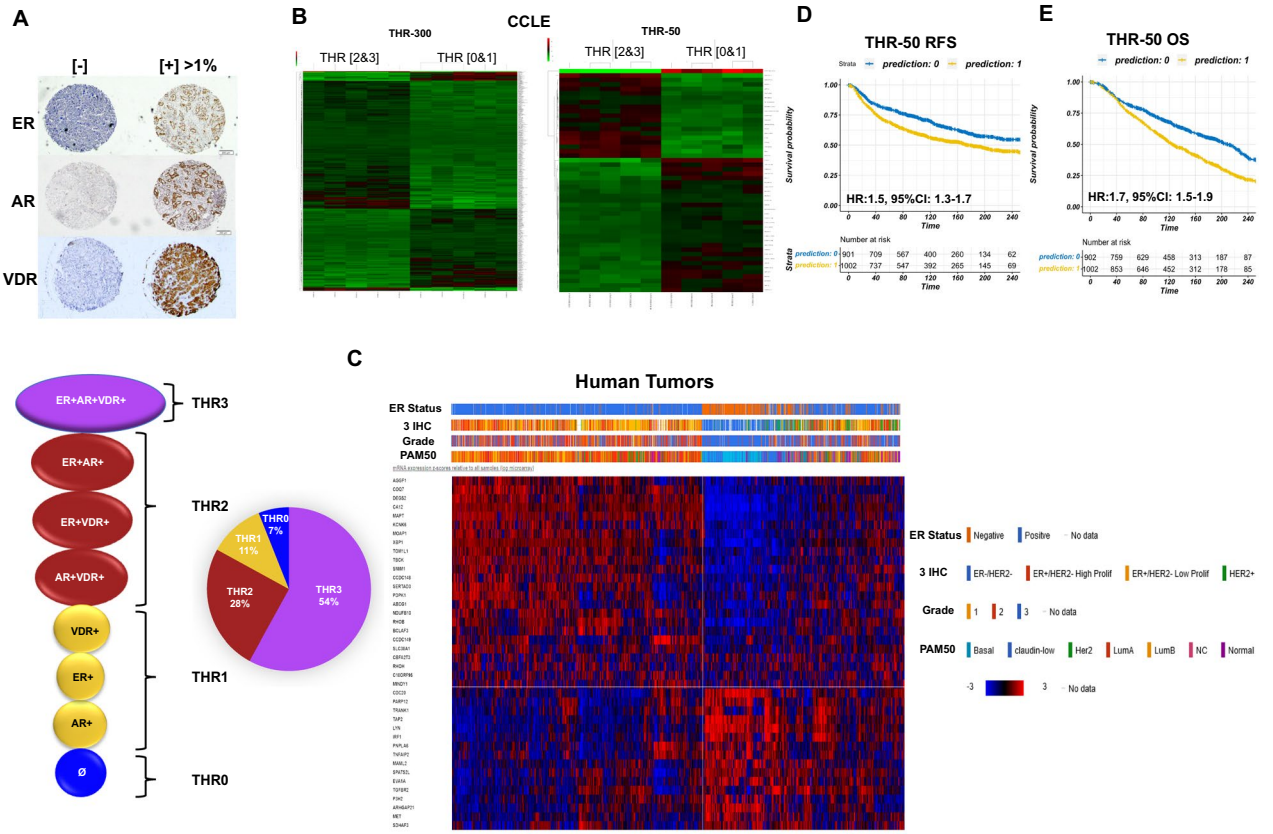


Fig. 1 Breast cancer classification based on triple-hormone receptor (THR) expression. **A** Immunohistochemical (IHC) staining with ER, AR, and VDR of tissue microarrays (TMAs) from breast cancer patients (top) identifies four distinct subtypes (bottom). Hormone receptor positive tumors were identified as those with > 1% protein expression. **B** Heatmaps showing the expression of the top 300 (left) and top 50 (right) differentially expressed genes between THR-[0/1] and THR-[2/3] cell lines in the Cancer Cell Line Encyclopedia (CCLE) dataset. **C** Heatmap showing the expression of the THR-50 genes in human samples from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort. The heatmap shows the samples annotation including ER status measured by IHC, clinical 3-gene classifier groups (based on ER, HER2, and MIB-1), histological grade, and PAM-50 groups. Red = high expression, blue = low expression. **D–E** Kaplan–Meier survival plots show the difference in recurrence-free survival (RFS) (**D**) and overall survival (OS) (**E**) between METABRIC samples predicted as 0 (low-risk) and 1 (high-risk) by THR-50. High-risk samples have significantly worse RFS (HR = 1.5, 95%CI: 1.3–1.7, $p < 0.0001$) and OS (HR = 1.7, 95%CI: 1.5–1.9, $p < 0.0001$) compared to low-risk samples. Survival time is in months. The hazard ratios and 95% confidence intervals are shown. THR: triple-hormone receptors; HR: hazard ratio; CI: confidence interval

ratio (HR) in multivariate analysis: THR-0 (HR = 6.9, CI: 3.3–14.3, n = 104); THR-1 (HR = 5.3, CI: 2.7–9.9, n = 185); THR-2 (HR = 2.9, CI: 1.6–5.2, n = 429); and THR-3 (HR = 1.0, n = 998) in the Nurses' Health Study dataset (n = 1716) [3].

Derivation of a triple hormone receptor mRNA signature: THR-50

Utilizing data from the CCLE dataset, we developed an mRNA signature that can distinguish between THR-0/1 and THR-2/3 BrCa [27]. We selected the top 50 most significant genes (lowest *p*-value) from the 600 differentially expressed mRNAs to investigate further, referred to as THR-50 signature hereafter, which allowed us to examine the THR-IHC index in publicly available BrCa gene expression datasets (Fig. 1B and Additional Files 2 and 4).

Validation of THR-50 in human breast *cancer*

Analysis of the METABRIC cohort (n = 1904) [28] demonstrates that the median expression of THR-50 divides BrCa into two major clusters (Fig. 1C). The genes associated with THR-[2/3] CCLE (Fig. 1B) are highly expressed in ER-positive BrCa within the METABRIC dataset (Fig. 1C), as expected. THR-50 is also significantly associated with RFS (HR = 1.5) and OS (HR = 1.7) (Fig. 1D–E). Additionally, even after adjusting for important variables such as age, tumor stage, and grade using a multivariate Cox proportional hazards model, THR-50 remains significantly associated with RFS and OS (Table 1),

Table 1 Multivariate survival analysis of THR-50

OS				
Variable		**HR**	**95% CI**	***p*-value**
THR-50 high-risk		1.4	1.2—1.6	2.4e-05
Age		1.03	1.02 – 1.04	< 2e-16
Tumor stage	Stage 2	1.5	1.3 – 1.8	4.97e-07
	Stage 3	2.9	2.2 – 3.8	7.49e-16
	Stage 4	4.7	2.3 – 9.7	2.02e-05
Histological Grade	Grade 2	1.1	0.8 – 1.4	0.72
	Grade 3	1.4	1.0 – 1.8	0.049
RFS				
Variable		**HR**	**95% CI**	***p*-value**
THR-50 high-risk		1.3	1.1 – 1.6	0.001
Age		0.99	0.985 – 0.998	0.02
Tumor stage	Stage 2	1.5	1.3 – 1.9	2.61e-05
	Stage 3	3.4	2.6 – 4.5	< 2e-16
	Stage 4	10.4	5.3 – 20.7	1.89e-11
Histological Grade	Grade 2	1.3	0.9 – 1.8	0.24
	Grade 3	1.6	1.1 – 2.3	0.02

Overall survival (OS) and Recurrence-free survival (RFS) in the METABRIC cohort using Cox proportional hazards model and including the THR-50-predicted risk groups, age, tumor stage, and histological grade. HR: Hazard ratio, 95% CI: 95% Confidence interval, *p*-value: Wald test *p*-value

underscoring the independent prognostic value of this signature.

Analysis of THR-50 across breast cancer subtypes

Next, we explored the association of THR-50 with survival outcomes across BrCa subtypes. Using the THR-50-derived risk score in the METABRIC cohort, we stratified patients into four equal groups with quartile 1 (Q1) and quartile 4 (Q4) representing the lowest and highest risk, respectively. We observed that patients in Q4 have significantly worse RFS compared to Q1 across ER-positive, ER-negative and HER2 + BrCa (Figure S2A, Additional File 5), as well as in the Luminal A (HR = 1.5) and Luminal B (HR = 2.4) BrCa subtypes (Figure S2B, Additional File 5). The distribution of THR-50 risk scores across different BrCa subtypes, defined by the PAM-50 and clinical 3-gene classification schemes, is illustrated in Figure S3, Additional File 5.

These findings were validated in the KMP cohort (n = 2,032, 50 studies) [29], where patients were stratified based on the average expression of THR-50, into low- and high-expression groups. We found that patients with low expression of THR-50 exhibit worse OS compared to those with high expression across ER + , AR + , ER-, Lum-A, Lum-B, HER2 + , and lymph-node positive (LN +) BrCa subtypes (Figure S4, Additional File 5). These results underscore the utility of the THR cell-of-origin signature in stratifying the risk of BrCa patients across diverse molecular and clinical subtypes.

THR-50 demonstrates promising performance relative to existing prognostic biomarker tests

Multigene biomarker tests such as Oncotype DX, PAM-50, MammaPrint, and EndoPredict are recommended by the American Society of Clinical Oncology (ASCO) for ER-positive, HER2-negative, and lymph node-negative BrCa [48–50]. However, they are not generally recommended by ASCO for ER-negative, HER2-positive, lymph node metastatic (> N1), or treated BrCa [51, 52].

In the KMP cohort (n = 2032), we evaluated the performance of THR-50 alongside PAM-50, MammaPrint, and Oncotype DX by categorizing patients based on the average expression of signature genes using optimal cut-offs. Remarkably, high average expression of THR-50 is significantly associated with better RFS compared to low expression in overall BrCa (HR = 2.04). Similarly, PAM-50 shows a significant association with RFS in overall BrCa (HR = 1.4), while MammaPrint and Oncotype DX do not ($p = 0.052$ and 0.13, respectively) (Figure S5, Additional File 5).

Next, we investigated the prognostic performance of these signatures across various BrCa groups. Noticeably, THR-50 demonstrates significant associations with RFS in multiple BrCa subgroups, including lymph node-positive (HR = 2.4), AR-positive (HR = 2.9), grade 2

(HR = 2.4), and grade 3 (HR = 1.6) BrCa (Fig. 2A). Similarly, PAM-50 demonstrates significant associations with RFS in the same groups, although with a slightly lower prognostic power compared to THR-50, except for grade 3 BrCa ($p = 0.29$) (Fig. 2A).

THR-50 identifies significant prognostic subgroups even within PAM-50 categories, revealing distinct RFS outcomes, in Luminal A (HR = 2.2), Luminal B (HR = 1.8), HER2-like (HR = 2.3), and basal-like BrCa (HR = 2.5) (Fig. 2B).

In contrast, MammaPrint demonstrates significance only in Luminal B (HR = 1.3) and HER2 + BrCa (HR = 1.5) subtypes. Oncotype DX, in comparison, shows significant associations with RFS across PAM-50 subtypes, but with reduced prognostic efficacy compared to THR-50, except for basal-like BrCa, where it did not reach significance ($p = 0.15$) (Fig. 2B).

These results indicate that THR-50 exhibits a significant prognostic value across diverse BrCa subtypes, unlike currently available tests.

Derivation and validation of THR-70

The results shown above using THR-50 suggest that the CCLE THR signature can be used to filter human tumor

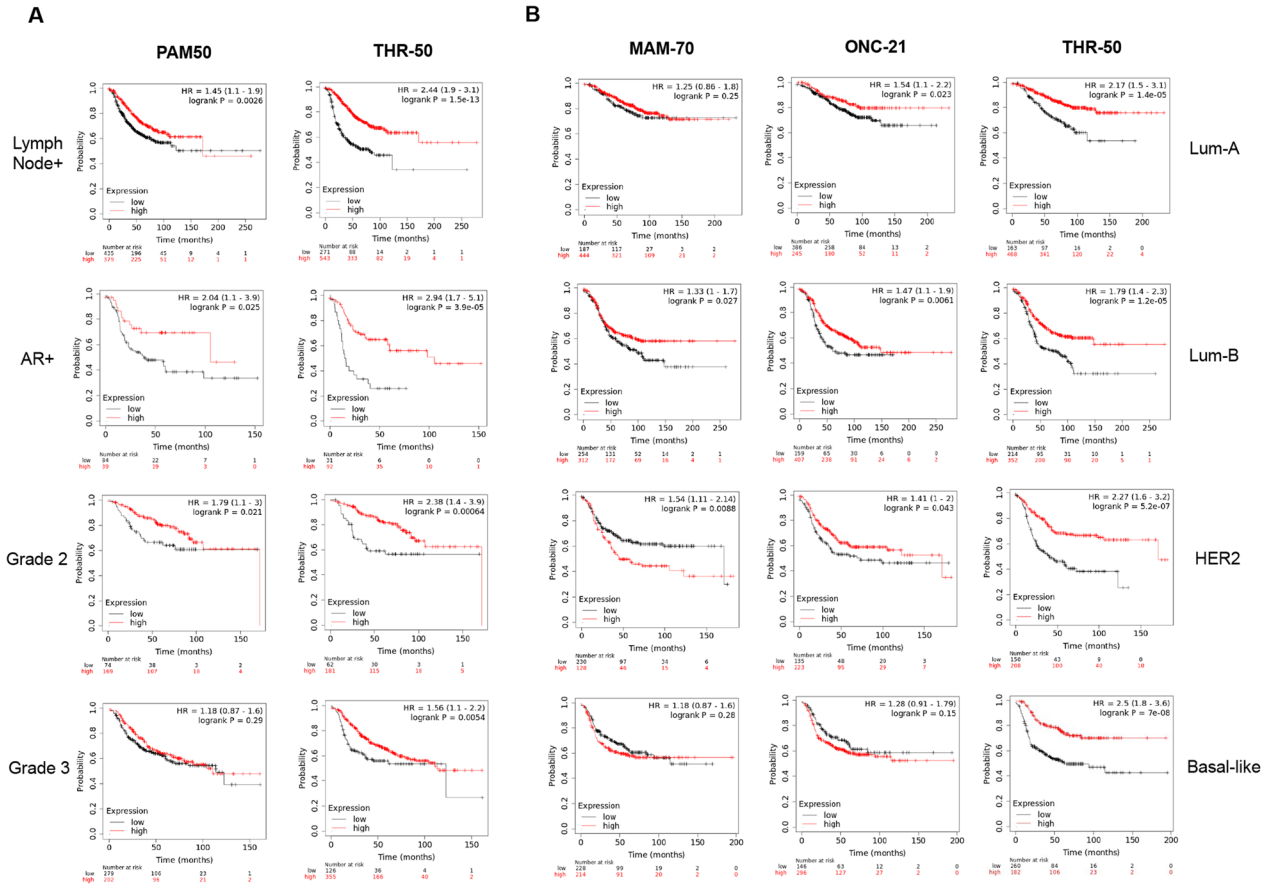


Fig. 2 THR-50 is significantly associated with recurrence-free survival (RFS) across different breast cancer clinical groups, outperforming established tests in the KMP cohort. **A** Kaplan–Meier survival plots comparing the prognostic power of THR-50 with PAM-50 using RFS in lymph-node positive, androgen receptor (AR) positive (AR+), grade 2 and grad 3 breast cancer. The analysis uses an independent validation cohort (KMP) comprising 2,032 samples from 50 gene expression datasets. Low (black line) and high (red line) expression groups are defined based on optimum cutoffs of the average expression levels of all signature genes. The reported *p*-values are derived from the log-rank test. The hazard ratios (HR) along with their corresponding 95% confidence intervals (CI) are shown. THR-50 results: lymph-node positive (HR = 2.4, CI:1.9–3.1, $p = 1.5e\text{-}13$), AR + (HR = 4.5, CI:2.0–11.0, $p = 0.0001$), Grade 2 (HR = 2.0, CI:1.1– 4.0, $p = 0.02$), Grade 3 (HR = 1.6, CI:1.1–2.5, $p = p = 0.02$). PAM-50 results: lymph-node positive (HR = 1.4, CI:1.1–1.9, $p = 0.0026$), AR + (HR = 2.0, CI:1.1–3.9, $p = 0.025$), Grade 2 (HR = 1.8, CI:1.1–3.0, $p = 0.021$), Grade 3 ($p = 0.29$). **B** Kaplan–Meier survival plots comparing the RFS between patients with low and high average expression of THR-50 genes across different PAM-50 groups: Lum-A (HR = 2.1, CI = 1.5–3.1, $p = 1.4e\text{-}05$), Lum-B (HR = 1.8, CI = 1.4–2.3, $p = 1.2e\text{-}05$), HER2-like (HR = 2.3, CI = 1.6–3.2, $p = 5.2e\text{-}07$), basal-like (HR = 2.5, CI = 1.8–3.6, $p = 7e\text{-}08$). The plot also shows Oncotype DX (ONC-21), and MammaPrint (MAM-70) HR, 95% CI, and *p-values*. CI: 95% confidence interval. HR: hazard ratio

gene expression data. Therefore, fine-tuning of the THR signature was conducted by overlapping the cell line expression profiles with human tumor tissue cohort, comprising 855 BrCa cases (BC855) [21, 32–34, 53].

Analysis of BC855 cohort reveals that THR categories differ from PAM-50 subtypes (Fig. 3A); demonstrating that each THR group encompasses diverse proportions of all six PAM-50 subtypes. For instance, the THR-1 cohort comprises Luminal A (15.7%), Luminal B (18.8%), HER2-enriched (18.5%), Claudin-low (13.9%), Normal-like (8.7%), and Basal-like (24.4%) PAM-50 clusters, indicating a balanced distribution (Fig. 3A). This distribution suggests that the poor outcomes in THR-1 are not attributed to a single PAM-50 cluster.

We found that 190 THR-associated mRNAs in cell lines (CCLE) overlap with THR signature in human BrCa tumors (BC855) (Fig. 3B). THR-70 refers to the top 70 genes identified by SAM fold change and *p*-value ranking among these 190 genes (Additional Files 3 and 4).

Next, by integrating THR-70 with the recently described single nucleus transcriptome data of healthy breast tissues [35], we show that THR-70 gene expression is enriched in normal human breast tissue. Interestingly, different THR-70 genes are enriched in proliferating (LASP-AP and LASP-BL) versus hormone sensing (LHS-HSα and LHS-HSβ) luminal epithelial breast cells (Fig. 3C).

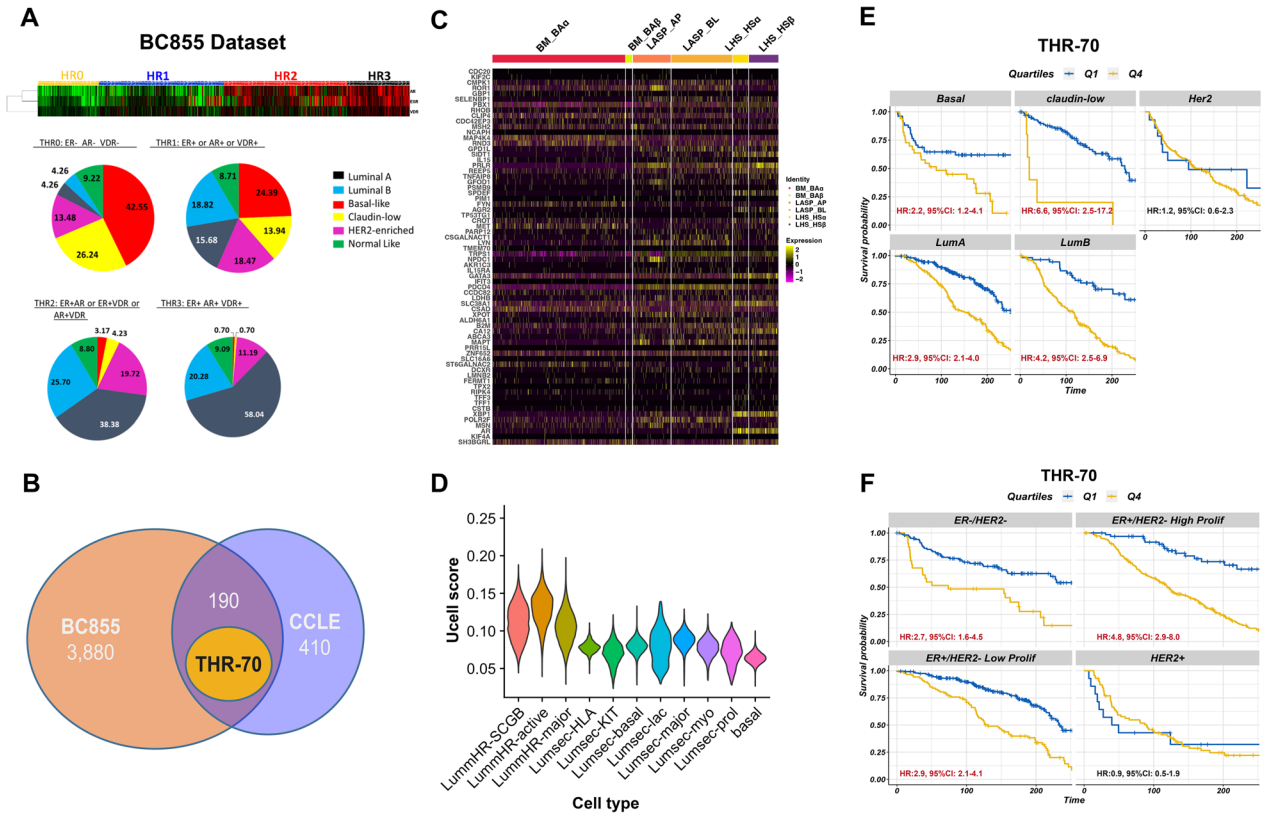


Fig. 3 Development and validation of THR-70. **A** Heatmap showing the expression of the three hormone receptors ER, AR, and VDR across the different triple hormone receptor (THR) groups (top). The Pie charts (bottom) show the percentages of PAM-50 subtypes in the different THR groups in the BC855 cohort. **B** Venn diagram showing the genes in common between the top differentially expressed genes (DEGs) between the THR-0/1 and THR-2/3 groups in the CCLE and BC855 cohorts, using $p < 0.05$ as a cut-off. The THR-70 signature comprises the top 70 DEGs in common between both cohorts based on SAM-fold expression. **C** Heatmap of the expression of THR-70 genes in normal breast epithelial clusters reported in Bhat-Nakshatri et al. Expression levels are z-score transformed. **D** Violin plots comparing the enrichment of THR-70 across normal breast epithelial clusters identified by Kumar et al. Signature scores (normalized U statistics between 0 and 1), shown on the Y axis, were computed using *UCell*. **E–F** Kaplan–Meier survival plots in the METABRIC cohort comparing the overall survival (OS) between patients predicted as low (Q1) and high-risk (Q4) by THR-70. The high-risk samples have significantly worse OS compared to low-risk samples in the PAM-50 basal (HR = 2.2, 95%CI: 1.2–4.1, $p = 0.01$), Claudin-low (HR = 6.6, 95%CI: 2.5–17.2), Luminal A (HR = 2.9, 95%CI: 2.1–4.0, $p < 0.0001$), and Luminal B (HR = 4.2, 95%CI: 2.5–6.9, $p < 0.0001$) (**E**). Additionally, in clinical 3-gene classifier ER-/HER2- (HR = 2.7, 95%CI: 1.6–4.5, $p < 0.0001$), ER + /HER2- high proliferation (HR = 4.8, 95%CI: 2.9–8.0, $p < 0.0001$), and ER + /HER2- low proliferation (HR = 2.9, 95%CI: 2.1–4.1, $p < 0.0001$) (**F**). Survival time is in months. Hazard ratios (HR) and 95% confidence intervals (CI) are shown. Statistically significant HRs are highlighted in red

Previously, we reported that the biomarker profiles of most human breast cancers, including the TNBC and basal-like cancers, are similar to normal luminal breast epithelium with the majority of tumors enriched for signatures derived from LHS and LASP cells [35, 54]. Consistent with this, we found that THR-70 is not enriched in normal basal-myoepithelial breast cells (BM_BAα and BM_BAβ) (Fig. 3C). Similar results are observed in the study by Kumar et al. [38], in which THR-70 is more enriched in luminal hormone-responsive cells (LummHR-SCGB, LummHR-active, and LummHR-major) compared to the luminal secretory and basal cells (Fig. 3D).

Having observed that THR-70 contains a breast cell-of-origin signature, we examined it in the KMP dataset revealing prognostic subgroups within PAM-50 categories including in Luminal A (HR = 1.6), Luminal B (HR = 2.0), HER2-like (HR = 2.1), and basal-like BrCa (HR = 2.2), as well as in lymph node-positive (HR = 2.8), AR-positive (HR = 1.6), grade 2 (HR = 2.5), and grade 3 (HR = 1.6) BrCa (Figure S6).

Next, we stratified patients within the METABRIC cohort based on THR-70, utilizing calculated risk scores, and observed that patients in the highest-risk category (Q4) have worse RFS and OS compared to those in the lowest-risk category (Q1) across all BrCa subtypes except HER2 + (Figs. 3E-F and S7-S8, Additional File 5).

To validate these findings in another dataset, we examined the THR-70 signature in the Meta-10 cohort, which comprises samples from ten different gene expression datasets [30] (Figures S9 and S10, Additional File 5), where THR-70 demonstrates significant association with RFS (HR = 2.5), distant metastasis-free survival (DMFS) (HR = 3.8), and survival in lymph node-positive (HR = 9.7), lymph node-negative (HR = 3.4), treated (HR = 4.4), and untreated (HR = 2.3) patients (Fig. 4).

Cumulatively, these results indicate that ASCO-recommended multigene biomarker tests exhibit varying prognostic efficacy across different BrCa subgroups. In contrast, THR signatures demonstrate consistent prognostic power across BrCa subtypes. These findings are notable as conventional prognostic signatures rarely

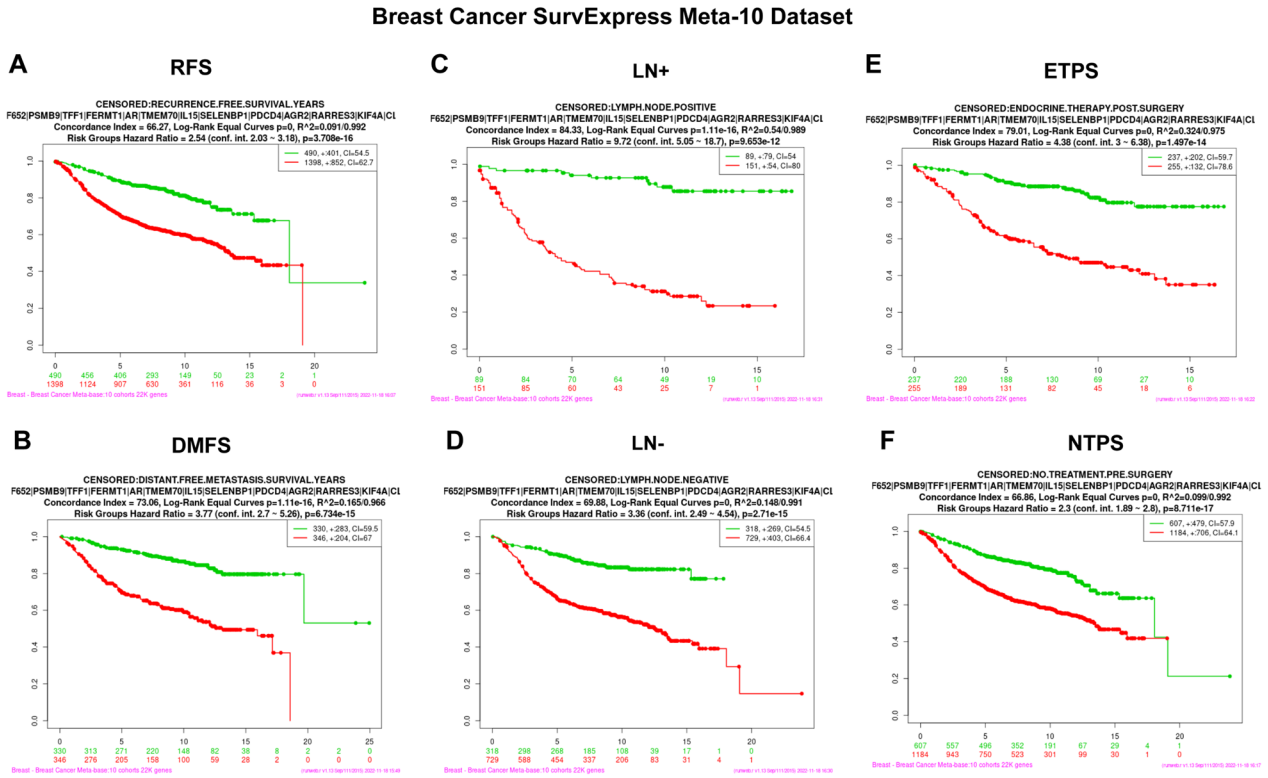


Fig. 4 THR-70 is prognostic in the SurvExpress Meta-10 cohort, comprising samples from 10 different datasets. THR-70 shows a significant association with recurrence-free survival (RFS, HR = 2.5 CI: 2.0—3.1, $p = 3.7e\text{-}16$) (**A**) and distant metastasis-free survival (DMFS, HR = 3.7 CI: 2.7—5.6, $p = 6.7e\text{-}15$) (**B**). It is prognostic in both lymph node positive (LN +, HR = 9.7 CI: 5.0—18.7, $p = 9.6e\text{-}12$) (**C**) and negative (LN-, HR = 3.3 CI: 2.4—4.5, $p = 2.7e\text{-}15$) (**D**) disease, as well as in patients treated with endocrine therapy post surgery (ETPS, HR = 4.3 CI: 3.0—6.3, $p = 1.4e\text{-}14$) (**E**), and those who did not receive neoadjuvant treatment (NTPS, HR = 2.3 CI: 1.8—2.8, $p = 8.7e\text{-}17$) (**F**). 95% confidence interval (CI); hazard ratio (HR)

correlate with such diverse aspects of tumor biology across multiple datasets (Figures S7-S10, Additional File 5), suggesting that a cell-of-origin signature may influence various facets of the tumor phenotype.

THR-50 and -70 are associated with hormone and immune gene set signatures

To gain deeper biological insights into the functional roles of THR-50 and THR-70, we conducted gene set variation analysis (GSVA). Consistent with their derivation based on hormone receptor protein expression, both THR signatures are enriched for AR and ER pathways, which are negatively correlated with PAM-50, MammaPrint, and a proliferation signature (PCNA-131) (Fig. 5A).

Traditional tissue-based prognostic signatures often reflect differences in tumor proliferation due to marked outcome differences between high-grade/high-proliferation tumors with poor outcomes and better-outcome tumors with low proliferation. It has been reported that proliferation-related genes are overrepresented in 88% of the BrCa prognostic signatures examined [55], and their removal substantially diminishes the prognostic efficacy across a majority of the 47 published signatures [56]. This suggests that many established BrCa prognostic tests may primarily operate as surrogate markers for proliferation [57]. Accordingly, the most notable positive associations observed in PAM-50 and MammaPrint through GSVA are with cell cycle and apoptosis pathways (Fig. 5A**).** In contrast, we aimed to reduce this proliferation bias by filtering THR signature from human tumors (BC855) with cell lines (CCLE-600) that exhibit uniformly high proliferation rates. Combining tumor and cell line data ensured that THR-70 is less influenced by proliferation effects (Fig. 5A), cell line artifacts, and non-tumor signals from tissue samples.

Notably, analysis of immune enrichment scores reveals that THR signatures are considerably associated with tumors with higher levels of central memory, gamma delta (γδ), CD4 + T, Th17, T follicular helper (Tfh), NK, and MAIT cell infiltrates (Fig. 5B). In contrast, PAM-50, MammaPrint, and PCNA-131 signatures are associated with tumors that are infiltrated with myeloid lineages such as neutrophils, dendritic

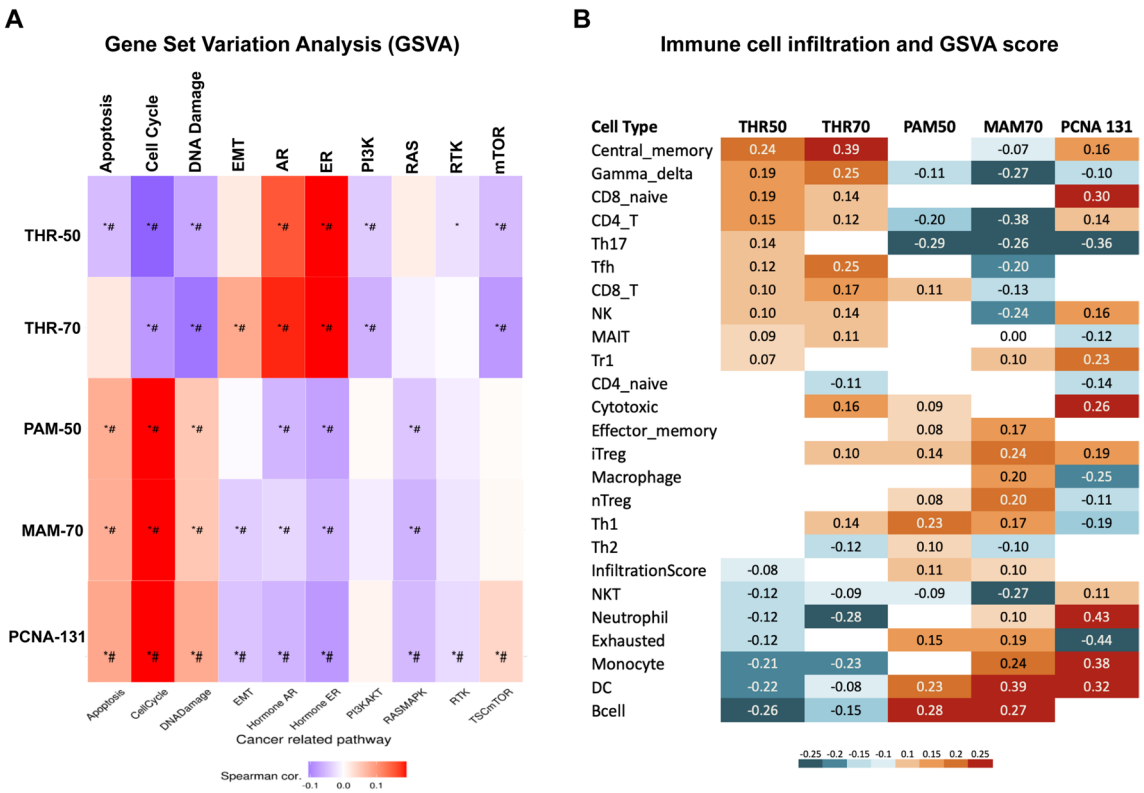


Fig. 5 The THR signatures are enriched in ER, AR, and immune pathways in gene set variation analysis (GSVA). **A** Heatmap showing the Spearman correlation coefficients between different breast cancer signatures (rows) and key cancer-associated pathways and biological processes (columns). Positive correlation (red), negative correlation (blue), *p*-value < 0.05 (*), false discovery rate (FDR) < 0.05 (#). **B** Immune cell type enrichment score heatmap (rows) in different breast cancer signatures (columns). Positive enrichment (red) and negative enrichment (blue)

cells, and monocytes (Fig. 5B). Some of these immune cells are known to influence BrCa outcomes, with CD8 + T cells generally associated with better outcomes and Treg/Th17 cells linked to poorer outcomes. Interestingly, γδT and Tfh are involved in anti-tumor cytotoxicity and antibody generation, respectively [58], suggesting that THR signatures may interact with both anti-tumor and pro-tumor immune infiltrates. Further research is needed to fully elucidate this immunological landscape, underscoring the intricate relationship between breast cell-of-origin and immune response in BrCa. In summary, these results indicate that THR signatures capture aspects of BrCa tumor biology that are not captured by standard prognostic tests.

Unsupervised clustering of breast cancer samples using THR signatures reveals distinct subtypes

We next explored the utility of the THR signature for de novo classification of BrCa. Using unsupervised clustering, we grouped samples in the METABRIC cohort based on their expression of THR-70, and subsequently analyzed the specific survival rates associated with each identified group. Our findings reveal that THR-70 divides BrCa into five distinct clusters: E1, E2a, E2b, E3, and PQNBC (Fig. 6A).

These clusters encompass a spectrum of ER-positive (E1-E3) clusters (Figure S11A, Additional File 5). Based on their overlapping survival curves, we consolidated E2a and E2b into a unified prognostic group named E2

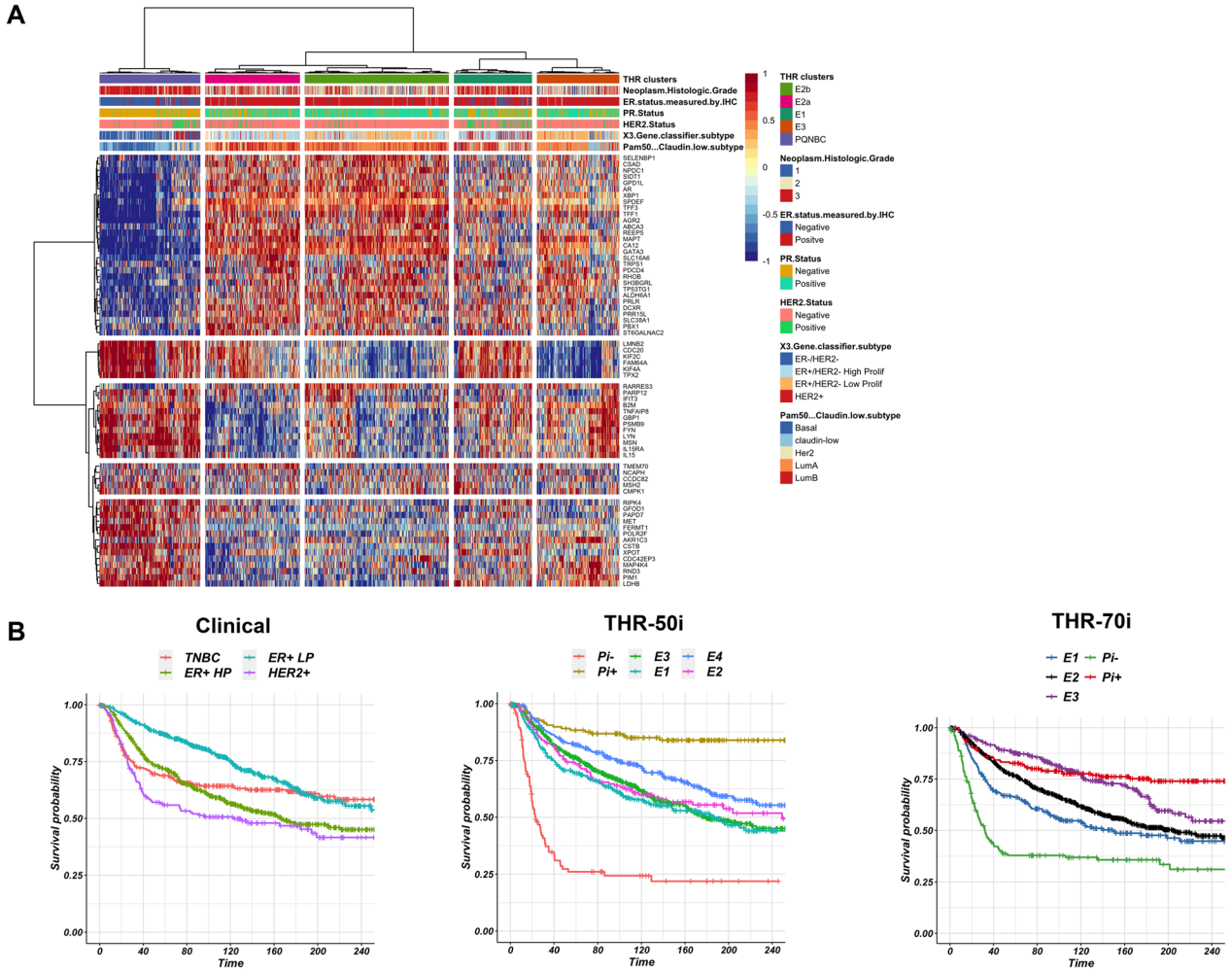


Fig. 6 Unsupervised clustering based on THR-70 uncovers five distinct breast cancer groups in the METABRIC cohort. **A** Heatmap showing the expression of THR-70 genes in the METABRIC cohort. Five distinct groups were identified: E1, E2a, E2b, E3, and PQNBC. **B** Kaplan–Meier survival plots comparing the 20 years recurrence-free survival (RFS) rates between different breast cancer groups identified by the by the 3-gene (ER, HER2, Mib-1 IHC) classifier (clinical), THR-50 combined with i20 (THR-50i), and THR-70 combined with i20 (THR-70i) signatures. THR: triple-hormone receptor. Pi + : pentaplex-negative (ER, PR, AR, VDR, and HER2), immune-positive tumors. Pi-: pentaplex-negative, immune-negative tumors. Survival time is in months

(Figure S11B-C, Additional File 5). The pentaplex-negative and quadruple-negative BrCa (PNBC and QNBC) clusters include breast tumors that are negative for (ER, PR, ±HER2, ±AR, and ±VDR) (Fig. 6A).

The THR heatmap clusters differ from existing categories of BrCa. For instance, while the THR-E clusters predominantly consist of Luminal A (Lum-A), Luminal B (Lum-B), and low-grade tumors, they also encompass other subtypes like HER2+, claudin-low, and high-grade tumors, albeit less frequently (Figure S12, Additional File 5).

Likewise, the PQNBC cluster contains multiple PAM-50 subtypes, including basal-like (49.7%), claudin-low (34.5%), and HER2-like (14.7%) (Figure S12, Additional File 5). These results suggest that the THR groups represent distinct classifications rather than merely renaming existing categories.

THR-70 clusters can be further stratified utilizing an immune signature

An immune signature consisting of 20 genes (referred to as i20, Additional File 4) was used to further divide the PQNBC cluster into two subgroups (PQNBC.i+and PQNBC.i-), with the PQNBC.i+subgroup characterized by higher immune infiltration than the PQNBC.i-subgroup. The combined THR-immune signatures are referred to as THR-50i and THR-70i.

Notably, when compared to the clinical 3-gene classification scheme, both THR-50i and THR-70i KM charts show significantly fewer survival curve crossovers (Fig. 6B). In contrast, ER-/HER2- tumors, typically considered poor outcome subtypes, exhibit survival curve overlap with ER-positive clusters (ER+LP/HP) (Fig. 6B). Similar findings were observed when comparing THR-70i to PAM-50, with THR-70i demonstrating clearer separation between different BrCa groups compared to PAM-50 (Fig. 7). These results align with previous studies showing multiple KM curve crossovers among PAM-50 subtypes [24, 59–64], highlighting that THR-immune signatures provide improved separation of outcome groups with reduced overlap.

As a cell-of-origin signature, THR genes exhibit mutations in less than 1–2% of BrCa cases. In contrast, PAM-50 signature genes have mutations in 35% of BrCa cases (Figure S13-14, Additional File 5). It has been proposed that early mutations might become redundant or non-essential as tumors progress, potentially altering the prognostic relevance of mutation-based signatures over time. Conversely, cell-of-origin signatures like THR may maintain stability throughout the tumor's lifespan.

Interestingly, i20 was also able to divide the ER-/HER2- group, defined by the clinical 3-gene classification scheme, into two distinct subgroups with significantly different survival rates (Figure S15B-C, Additional File 5), suggesting that i20 immune signature can further enhance the granularity of existing classifiers.

Stratification of ER+, TNBC, and HER2 cancer with THR.i signature

In ER-negative BrCa, we observed a 1.5-fold difference in survival probability using the clinical 3-gene (ER/HER2/Mib-1) and PAM-50 classification methods. In contrast, there is a 15-fold difference in survival probability between predominantly ER-negative PQNBC.i+and PQNBC.i- cohorts (HR=15.7, 95%CI: 8.5–29.0). Therefore, compared to the 3-gene classifier and PAM-50, the THR-70i signature demonstrates a ten-fold improvement in distinguishing ER-negative breast tumor subtypes with markedly poor and favorable outcomes (Fig. 7A).

For ER-positive BrCa, while the 3-gene and PAM-50 methods identify two ER-positive subtypes with hazard ratios differing by 1.5- to 1.8-fold, THR-70 delineates three distinct ER-positive clusters (E1, E2, and E3) with a survival range differing by 2.1-fold (Fig. 7B).

Combining cell-of-origin, immune, and genetic biomarkers

We previously demonstrated that HER2-amplified BrCa does not align with a specific cell subtype in normal human breast. Consistent with their pathogenesis, HER2+cancers exhibit marker profiles spanning various normal luminal cell types [3]. Therefore, we did not anticipate that HER2+tumors would form a distinct cluster based on cell-of-origin signatures, which indeed was observed in THR cluster heatmaps (Fig. 6). Thus, to further stratify BrCa we coupled THR-70i with HER2+(THR-70Hi) that identified six BrCa groups with different survival estimates: PQNBC.i-, PQNBC.i+, E1, E2, E3, and HER2+(Figs. 8A and S15A-C, Additional File 5). The PQNBC subtype includes triple-negative (ER, PR, HER2) breast cancers (TNBC) that may lack either AR or VDR (quadruple-negative) and both AR and VDR (pentaplex-negative).

The survival curves of THR-70Hi groups generally do not overlap, demonstrating survival differences ranging up to 5.8-fold in both univariate (Fig. 8A) and multivariate (Fig. 8C) survival analyses. In comparison, the PAM-50 clusters exhibit up to a 3.6-fold survival difference range among its groups (Fig. 8B). However, the basal-like subtype's survival curve crosses over HER2, Luminal B, and Luminal A subtypes around 5-, 10-, and 20 years, respectively, complicating its assessment (Fig. 8B).

Next, we examined the relationship between THR and HER2 in more detail. First, we note that HER2 status has no effect on RFS in the THR-70 PQNBC cluster ($p=0.28$; Figure S16A, Additional File 5), emphasizing the dominance of the immune signature in this group.

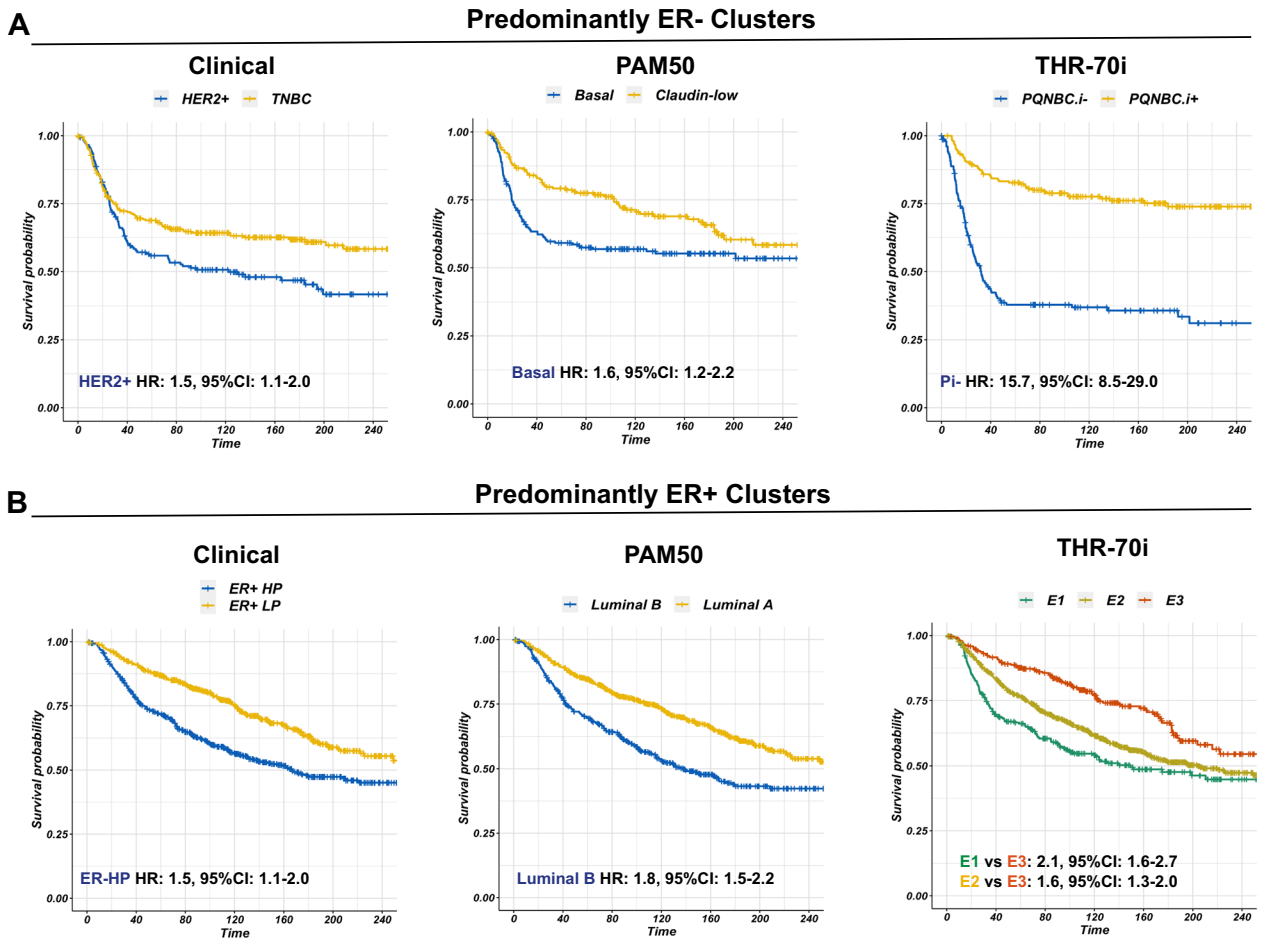
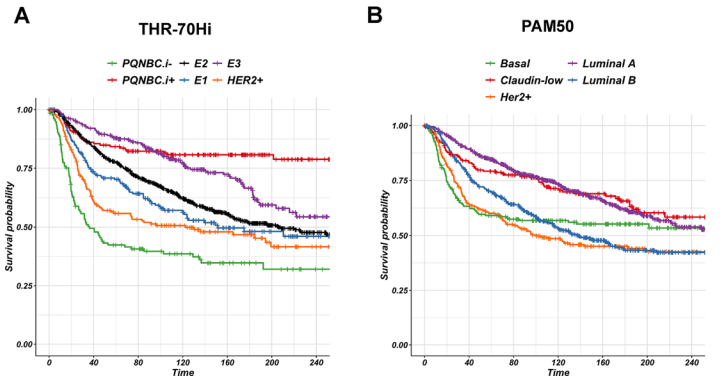


Fig. 7 THR-70 improves the identification of ER-negative (ER-) and ER-positive (ER+) subgroups with distinct survival rates compared to the clinical 3-gene classifier and PAM-50. **A** Kaplan–Meier (KM) survival plots comparing the 20-year recurrence-free survival (RFS) in ER-negative breast cancer groups identified by clinical 3-gene classifier: HER2 + HR = 1.5, 95%CI: 1.1–2.0, $p = 0.001$ vs. TNBC (left panel); PAM-50 classifier: basal HR = 1.6, 95%CI: 1.2–2.2, $p = 0.01$ vs. claudin-low (middle panel), and THR-70i: PQNBC.i- HR = 15.7, 95%CI: 8.5–29.0, $p < 0.0001$ vs. PQNBC.i + (right panel). **B** KM plots comparing the 20-year RFS in ER + groups identified by clinical 3-gene classifier: ER + HP HR = 1.7, 95%CI: 1.4–2.1, $p < 0.0001$ vs. ER + LP (left panel), PAM-50: luminal B HR = 1.8, 95%CI: 1.5–2.2, $p < 0.0001$ vs. Luminal A (middle panel), and THR-70i: E1 HR = 2.1, 95%CI: 1.6–2.7, $p < p < 0.0001$; E2 HR = 1.6, 95%CI: 1.3–2.0, $p = 0.0001$, VS. E3 (right panel). Survival time is in months. The hazard ratios (HR) and 95% confidence intervals (CI) are shown. HP: high proliferation, LP: low proliferation

Interestingly, we observed that HER2 status correlates with RFS in ER-positive THR-E1 ($p = 0.012$) and THR-E2 ($p = 0.001$) clusters but not in the THR-E3 cluster (Figure S17, Additional File 5). This suggests that HER2 status influences outcomes more prominently in THR-low clusters (E1/2) compared to THR-high E3. If validated, this finding could aid in stratifying ER + /HER2 + patients; those patients with HER2+ tumors that are triple-positive for ER, AR, and VDR may experience better outcomes than HER2 + patients with ER-positive but VDR and/or AR-negative BrCa. Additionally, we found that THR-50 can also stratify HER2 + patients into two distinct survival groups (HR = 2.2, CI: 1.4–3.7) (Figure S16C, Additional File 5). In summary, the THR cell-of-origin signature unveils a novel aspect of BrCa biology divergent from existing clustering methods, offering a robust foundation for integrating other prognostic biomarkers to enhance BrCa stratification.

Discussion

It is increasingly recognized that the cell-of-origin significantly impacts tumor biology and response to treatment. Therefore, understanding the lineage/differentiation state from which the tumor arises, or toward which it differentiates, provides crucial insights into its prognosis and potential therapeutic strategies. This principle is exemplified by the classification scheme of hematological malignancies, which uses stable reference points provided by



C

Multivariate Survival Analysis

THR-70Hi			PAM-50		
	HR	95% CI		HR	95% CI
PQNBC.i+	Reference		Claudin-low	Reference	
E3	1.6	1.1 – 2.4	Lum-A	0.98	0.7 – 1.3
E2	2.5	1.7 – 3.6			
E1	3.0	2.0 – 4.5	Lum-B	1.7	1.3 – 2.3
HER2	3.7	2.5 – 5.6	HER2-like	2.0	1.4 – 2.7
PQNBC.i-	5.8	3.9 – 8.7	Basal-like	1.7	1.2 – 2.3

Fig. 8 THR-70 coupled with immune signature (i20) and HER2 (THR-70Hi) captures more granular breast cancer groups compared to PAM-50. **A** Kaplan–Meier (KM) survival chart shows 20-year recurrence-free survival (RFS) of different patient subgroups identified by THR-70, i20, and HER2 classifier (THR70-Hi): E3 (purple), E2 (black), E1 (blue), HER2 + (yellow), PQNBC.i- (green) and PQNBC.i + (red). PNBC subtype includes breast cancers that are negative for ER, PR, HER2, AR and VDR. QNBC subtype includes breast cancers that are negative for ER, PR, HER2, ∓ AR or VDR. **B** KM survival chart shows 20-year RFS of different patient subgroups identified by PAM-50: Lum-A (purple), Lum-B (blue), HER2-like (yellow), basal-like (green), and claudin-low (red). **C** Multivariate analysis of Hazard ratios (HR) and 95% confidence interval (95% CI) for RFS by THR-70Hi and PAM-50 breast cancer groups using a Cox proportional hazards model

normal hematological cell types. Building on this concept, we have previously identified distinct differentiation types in normal breast luminal epithelial (NBLE) cells based on the expression of ER, AR, and VDR and DNA methylation profiles [3, 5, 6]. We have previously demonstrated that drugs targeting AR and VDR exhibit an additive effect in reducing proliferation in TNBC expressing both receptors. Additionally, our research has shown that combining AR and VDR hormone treatments with chemotherapy also yields additive effects in TNBC [3, 65]. These findings underscore the potential predictive utility of the THR signature in guiding therapeutic decisions for patients with TNBC.

In this study, we introduced the THR signature as a novel BrCa classification system based on cell lineage. One of the key strengths of our study is the utilization of large and well-characterized cohorts of BrCa, which allowed us to validate the prognostic performance of the THR signatures across various clinical and molecular subgroups. Furthermore, by comparing the performance of these signatures with existing classifiers, we demonstrated their significant value in BrCa prognostication. By integrating the cell-of-origin signature with existing molecular classifications and a novel immune signature (i20) we demonstrate robust prognostic value across various breast cancer subtypes, revealing more pronounced survival differences compared to current clinical and molecular subtyping methods. We also note that THR is not dominated by a proliferation signal, unlike most BrCa prognostic signatures.

Our findings with THR signatures enable a structured classification system similar to the taxonomy of species, categorizing tumors systematically based on hierarchical levels: phylum (organ), class (tissue), order (cell type), genus (differentiation state), and species (genetic/epigenetic alterations). For instance, BrCa can be classified as breast (phylum), lobule (class), duct (class), epithelium (order), THR status (genus), and HER2 status (species). Immunological markers like i20 can further describe the tumor's ecosystem. This modular method can eventually stratify BrCa patients by incorporating additional tumor hallmarks such as angiogenesis, proliferation, apoptosis, senescence, immunity, and invasion [66, 67].

In contrast with this step-wise taxonomic approach, with a single classifier for each hierarchical level, Oncotype DX [23], MammaPrint [21, 68] and Prosigna (PAM-50) [24] employ hybrid taxa composed of both cell type markers (*ER, PR, KRT5, KRT14,* and *KRT17*) and genetic markers (*MYC, EGRF, GRB7, MDM2, FGFR,* and *HER2*). While such composite signatures appear to be useful in clinical practice, deconstructing them retroactively and assigning biological meaning is difficult [69], which may explain why they have not evolved in tandem with our growing understanding of BrCa pathophysiology [70].

A recent study titled "Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer" convincingly illustrates the importance of cell-of-origin [71]. However, since these cell specific patterns can involve over thirty percent of the entire epigenome [72, 73], these observations have been difficult to translate into clinical tools [74–83]. Our research shows how these genome-wide patterns can be translated into practical signatures that predict overall survival and recurrence-free survival in BrCa patients. The THR signature provides a more comprehensive and clinically

meaningful prognostic model for BrCa by incorporating breast cell-of-origin information and THR status. Future research should focus on validating these signatures in clinical trials and investigating their utility as predictive biomarkers for treatment response and therapeutic target discovery.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13058-024-01876-9.

Additional file 1. Supplementary Material and Methods. Detailed description of the methods implemented in our study. This includes the derivation of the THR signatures, survival analysis, comparison with other existing signatures, and gene set enrichment analysis.

Additional file 2. Derivation of THR-50. THR-50 was derived by comparing the expression profiles of THR-0/1 and THR-2/3 breast cancer cell lines in the cancer cell line encyclopedia (CCLE) dataset.

Additional file 3. Derivation of THR-70. THR-70 was derived by overlapping the differentially expressed genes from the breast cancer cell lines (CCLE cohort) and human tissue samples (BC855 cohort).

Additional file 4. Gene symbols for the THR-50, THR-70, i20, Oncotype DX, PAM-50, and MammaPrint signatures

Additional file 5. Supplementary Figures.

Additional file 6. Raw image files for Figs. 1B-C, 2, and 3.

Additional file 7. Raw image files for Figs. 4, 5, and 6.

Additional file 8. Raw image files for Figs. S4–S6, S9–10, and S13–14.

Acknowledgements
The funder did not play a role in the design of the study; the collection, analysis, and interpretation of the data; the writing of the manuscript; and the decision to submit the manuscript for publication.

Author contributions
T.I. formulated the research question and study design. T.I., M.O., and C.E. performed the analysis. T.I. and M.O. wrote the manuscript draft. C.H., R.T., H.N., and L.M. edited the manuscript. All authors have read and agreed to the final version of the manuscript.

Funding
Weill Cornell Medicine Institutional support.

Availability of data and materials
The expression profiles of the breast cancer cell lines used to develop the THR-50 signature were derived from Cancer Cell Line Encyclopedia (CCLE) dataset which is publicly available through the Gene Expression Omnibus (GEO) under the accession code GSE36133. The METABRIC and TCGA breast cancer datasets used in this study are publicly available through cBioPortal using the following links: https://www.cbioportal.org/study/summary?id=brca_metabric and https://www.cbioportal.org/study/summary?id=brca_tcga_pub2015. The code used to perform the analysis can be accessed using the following GitHub repository: https://github.com/MohamedOmar2020/BreastCancer_THRsignature. The raw materials and images from KMP and SurvExpress are provided in Additional Files 6–8.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA. [2]Dana-Farber Cancer Institute, Boston, MA, USA. [3]Department of Pathology, Virginia Commonwealth University, Richmond, VA 23298, USA. [4]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA. [5]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA. [6]Departments of Surgery, Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN 46202, USA. [7]New York-Presbyterian, Brooklyn Methodist Hospital, New York, NY, USA.

Received: 14 February 2024 Accepted: 29 July 2024
Published online: 13 September 2024

References

1. Khoury JD, Solary E, Abla O, Akkari Y, Alaggio R, Apperley JF, Bejar R, Berti E, Busque L, Chan JK, Chen W. The 5th edition of the World Health Organization classification of haematolymphoid tumours: myeloid and histiocytic/dendritic neoplasms. Leukemia. 2022;36(7):1703–19.
2. Alaggio R, Amador C, Anagnostopoulos I, Attygalle AD, de Araujo IB, et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. Leukemia. 2022;36(7):1720–48.
3. Santagata S, Thakkar A, Ergonul A, Wang B, Woo T, Hu R, et al. Taxonomy of breast cancer based on normal cell phenotype predicts outcome. J Clin Invest. 2014;124(2):859–70.
4. Santagata S, Ince TA. Normal cell phenotypes of breast epithelial cells provide the foundation of a breast cancer taxonomy. Expert Rev Anticancer Ther. 2014;14(12):1385–9.
5. Houseman EA, Kile ML, Christiani DC, Ince TA, Kelsey KT, Marsit CJ. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. BMC Bioinform. 2016;17(1):259.
6. Houseman EA, Ince TA. Normal cell-type epigenetics and breast cancer classification: a case study of cell mixture-adjusted analysis of DNA methylation data from tumors. Cancer Inform. 2014;13(Suppl 4):53–64.
7. Dontu G, Ince TA. Of mice and women: a comparative tissue biology perspective of breast stem cells and differentiation. J Mammary Gland Biol Neoplasia. 2015;20(1–2):51–62.
8. Allison KH, Hammond MEH, Dowsett M, McKernin SE, Carey LA, Fitzgibbons PL, et al. Estrogen and progesterone receptor testing in breast cancer: ASCO/CAP guideline update. JCO. 2020;38(12):1346–66.
9. Freake HC, Abeyasekera G, Iwasaki J, Marcocci C, MacIntyre I, McClelland RA, et al. Measurement of 1,25-dihydroxyvitamin D3 receptors in breast cancer and their relationship to biochemical and clinical indices. Cancer Res. 1984;44(4):1677–81.
10. Al-Azhri J, Zhang Y, Bshara W, Zirpoli G, McCann SE, Khoury T, et al. Tumor expression of vitamin D receptor and breast cancer histopathological characteristics and prognosis. Clin Cancer Res. 2017;23(1):97–103.
11. Huss L, Butt ST, Borgquist S, Elebro K, Sandsveden M, Manjer J, et al. Levels of vitamin D and expression of the vitamin D receptor in relation to breast cancer risk and survival. Nutrients. 2022;14(16):3353.
12. Huss L, Butt ST, Borgquist S, Elebro K, Sandsveden M, Rosendahl A, et al. Vitamin D receptor expression in invasive breast tumors and breast cancer survival. Breast Cancer Res. 2019;21(1):84.
13. Narayanan R, Dalton JT. Androgen receptor: a complex therapeutic target for breast cancer. Cancers (Basel). 2016;8(12):108.
14. Ricciardelli C, Bianco-Miotto T, Jindal S, Butler LM, Leung S, McNeil CM, et al. The magnitude of androgen receptor positivity in breast cancer is critical for reliable prediction of disease outcome. Clin Cancer Res. 2018;24(10):2328–41.
15. Wang C, Pan B, Zhu H, Zhou Y, Mao F, Lin Y, et al. Prognostic value of androgen receptor in triple negative breast cancer: A meta-analysis. Oncotarget. 2016;7(29):46482–91.
16. Kensler KH, Poole EM, Heng YJ, Collins LC, Glass B, Beck AH, et al. Androgen receptor expression and breast cancer survival: results from the nurses' health studies. JNCI J Natl Cancer Inst. 2019;111(7):700–8.

17. Hu R, Dawood S, Holmes MD, Collins LC, Schnitt SJ, Cole K, et al. Androgen receptor expression and breast cancer survival in postmenopausal women. Clin Cancer Res. 2011;17(7):1867–74.
18. Bhattacharya U, Kamran M, Manai M, Cristofanilli M, Ince TA. Cell-of-origin targeted drug repurposing for triple-negative and inflammatory breast carcinoma with HDAC and HSP90 inhibitors combined with niclosamide. Cancers (Basel). 2023;15(2):332.
19. Cardoso F, Paluch-Shimon S, Senkus E, Curigliano G, Aapro MS, André F, et al. 5th ESO-ESMO international consensus guidelines for advanced breast cancer (ABC 5). Ann Oncol. 2020;31(12):1623–49.
20. JohnFR R, Paridaens RJ, Lichfield J, Bradbury I, Campbell C. Meta-analyses of phase 3 randomised controlled trials of third generation aromatase inhibitors versus tamoxifen as first-line endocrine therapy in postmenopausal women with hormone receptor-positive advanced breast cancer. Eur J Cancer. 2021;145:19–28.
21. van de Vijver MJ, He YD, et al. A gene-expression signature as a predictor of survival in breast cancer. New England J Med. 2002;347(25):1999–2009.
22. Buyse M, Loi S, et al. Validation and clinical utility of a 70-Gene prognostic signature for women with node-negative breast cancer. JNCI J Natl Cancer Inst. 2006;98(17):1183–92.
23. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A Multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004;351(27):2817–26.
24. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160–7.
25. Harris LN, Ismaila N, McShane LM, Andre F, Collyar DE, Gonzalez-Angulo AM, et al. Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: american society of clinical oncology clinical practice guideline. J Clin Oncol. 2016;34(10):1134–50.
26. Duffy MJ, Harbeck N, Nap M, Molina R, Nicolini A, Senkus E, et al. Clinical use of biomarkers in breast cancer: Updated guidelines from the European Group on Tumor Markers (EGTM). Eur J Cancer. 2017;1(75):284–98.
27. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603–7.
28. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346–52.
29. Győrffy B. Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. Comput Struct Biotechnol J. 2021;1(19):4101–9.
30. Aguirre-Gamboa R, Gomez-Rueda H, Martínez-Ledesma E, Martínez-Torteya A, Chacolla-Huaringa R, Rodriguez-Barrientos A, et al. SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. PLoS ONE. 2013;8(9): e74250.
31. Ivshina AV, George J, Senko O, Mow B, Putti TC, Smeds J, et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Res. 2006;66(21):10292–301.
32. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet. 2005;365(9460):671–9.
33. Bos PD, Zhang XHF, Nadal C, Shu W, Gomis RR, Nguyen DX, et al. Genes that mediate breast cancer metastasis to the brain. Nature. 2009;459(7249):1005–9.
34. Minn AJ, Gupta GP, Siegel PM, Bos PD, Shu W, Giri DD, et al. Genes that mediate breast cancer metastasis to lung. Nature. 2005;436(7050):518–24.
35. Bhat-Nakshatri P, Khatpe AS, Chen D, Batic K, Mang H, Herodotou C, et al. Signaling pathway alterations driven by BRCA1 and BRCA2 germline mutations are sufficient to initiate breast tumorigenesis by the PIK3CAH1047R oncogene. Cancer Res Commun. 2024;4(1):38–54.
36. Gray GK, Li CMC, Rosenbluth JM, Selfors LM, Girnius N, Lin JR, et al. A human breast atlas integrating single-cell proteomics and transcriptomics. Dev Cell. 2022;57(11):1400-1420.e7.
37. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888-1902.e21.
38. Kumar T, Nee K, Wei R, He S, Nguyen QH, Bai S, et al. A spatially resolved single-cell genomic atlas of the adult human breast. Nature. 2023;620(7972):181–91.
39. Andreatta M, Carmona SJ. UCell: Robust and scalable single-cell gene signature scoring. Comput Struct Biotechnol J. 2021;1(19):3796–8.
40. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep. 1966;50(3):163–70.
41. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. J Royal Stat Soc Ser A (General). 1972;135(2):185–207.
42. Liu CJ, Hu FF, Xia MX, Han L, Zhang Q, Guo AY. GSCALite: a web server for gene set cancer analysis. Bioinformatics. 2018;34(21):3771–2.
43. Miao YR, Zhang Q, Lei Q, Luo M, Xie GY, Wang H, et al. ImmuCellAI: a unique method for comprehensive T-cell subsets abundance prediction and its application in cancer immunotherapy. Adv Sci. 2020;7(7):1902880.
44. Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963;58(301):236–44.
45. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model [Internet]. New York, NY: Springer; 2000 [cited 2023 Feb 20]. (Dietz K, Gail M, Krickeberg K, Samet J, Tsiatis A, editors. Statistics for Biology and Health). https://doi.org/10.1007/978-1-4757-3294-8
46. Cox DR. Regression models and life-tables. J Roy Stat Soc: Ser B (Methodol). 1972;34(2):187–220.
47. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for generalized linear models via coordinate descent. J Stat Softw. 2010;2(33):1–22.
48. Vieira AF, Schmitt F. An update on breast cancer multigene prognostic tests—emergent clinical biomarkers. Front Med. 2018. https://doi.org/10.3389/fmed.2018.00248.
49. Barbi M, Makower D, Sparano JA. The clinical utility of gene expression assays in breast cancer patients with 0–3 involved lymph nodes. Ther Adv Med Oncol. 2021;14(13):17588359211038468.
50. The Way of the Future: Personalizing Treatment Plans Through Technology | American Society of Clinical Oncology Educational Book [Internet]. [cited 2023 Jun 27]. https://doi.org/10.1200/EDBK_320593?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed
51. Varga Z, Sinn P, Fritzsche F, von Hochstetter A, Noske A, Schraml P, et al. Comparison of EndoPredict and Oncotype DX test results in hormone receptor positive invasive breast cancer. PLoS ONE. 2013;8(3): e58483.
52. Bösl A, Spitzmüller A, Jasarevic Z, Rauch S, Jäger S, Offner F. MammaPrint versus EndoPredict: Poor correlation in disease recurrence risk classification of hormone receptor positive breast cancer. PLoS ONE. 2017;12(8): e0183458.
53. Adamo B, Deal AM, Burrows E, Geradts J, Hamilton E, Blackwell KL, et al. Phosphatidylinositol 3-kinase pathway activation in breast cancer brain metastases. Breast Cancer Res. 2011;13(6):R125.
54. Bhat-Nakshatri P, Gao H, Khatpe AS, et al. Single-nucleus chromatin accessibility and transcriptomic map of breast tissues of women of diverse genetic ancestry. Nat Med. 2024. https://doi.org/10.1038/s41591-024-03011-9
55. Solé X, Bonifaci N, López-Bigas N, Berenguer A, Hernández P, Reina O, et al. Biological convergence of cancer signatures. PLoS ONE. 2009;4(2): e4544.
56. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput Biol. 2011;7(10): e1002240.
57. Nunes AT, Collyar DE, Harris LN. Gene expression assays for early-stage hormone receptor-positive breast cancer: understanding the differences. JNCI Cancer Spectr. 2017;1(1):pkx008.
58. Byrne A, Savas P, Sant S, Li R, Virassamy B, Luen SJ, et al. Tissue-resident memory T cells in breast cancer control and immunotherapy responses. Nat Rev Clin Oncol. 2020;17(6):341–8.
59. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci. 2001;98(19):10869–74.
60. Perou CM. Molecular stratification of triple-negative breast cancers. Oncologist. 2010;15(S5):39–48.
61. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res. 2010;12(5):R68.
62. Milioli HH, Vimieiro R, Tishchenko I, Riveros C, Berretta R, Moscato P. Iteratively refining breast cancer intrinsic subtypes in the METABRIC dataset. BioData Mining. 2016;9(1):2.
63. Jiang YZ, Ma D, Suo C, Shi J, Xue M, Hu X, et al. Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies. Cancer Cell. 2019;35(3):428-440.e5.

64. Vallon-Christersson J, Häkkinen J, Hegardt C, Saal LH, Larsson C, Ehinger A, et al. Cross comparison and prognostic assessment of breast cancer multigene signatures in a large population-based contemporary clinical series. Sci Rep. 2019;9(1):12184.
65. Thakkar A, Wang B, Picon-Ruiz M, Buchwald P, Ince TA. Vitamin D and androgen receptor-targeted therapy for triple-negative breast cancer. Breast Cancer Res Treat. 2016;157(1):77–90.
66. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57–70.
67. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–74.
68. Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, Van Der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415(6871):530–6.
69. Emmert-Streib F, Manjang K, Dehmer M, Yli-Harja O, Auvinen A. Are there limits in explainability of prognostic biomarkers? Scrutinizing biological utility of established signatures. Cancers. 2021;13(20):5087.
70. Manjang K, Tripathi S, Yli-Harja O, Dehmer M, Glazko G, Emmert-Streib F. Prognostic gene expression signatures of breast cancer are lacking a sensible biological meaning. Sci Rep. 2021;11(1):156.
71. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell. 2018;173(2):291-304.e6.
72. Mancarella D, Plass C. Epigenetic signatures in cancer: proper controls, current challenges and the potential for clinical translation. Genome Med. 2021;13(1):23.
73. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nat Rev Genet. 2010;11(7):476–86.
74. Xin L. Cells of origin for cancer: an updated view from prostate cancer. Oncogene. 2013;32(32):3655–63.
75. Merritt MA, Bentink S, Schwede M, Iwanicki MP, Quackenbush J, Woo T, et al. Gene expression signature of normal cell-of-origin predicts ovarian tumor outcomes. PLoS ONE. 2013;8(11): e80314.
76. Bhagirath D, Zhao X, West WW, Qiu F, Band H, Band V. Cell type of origin as well as genetic alterations contribute to breast cancer phenotypes. Oncotarget. 2015;6(11):9018–30.
77. Bu W, Liu Z, Jiang W, Nagi C, Huang S, Edwards DP, et al. Mammary precancerous stem and non-stem cells evolve into cancers of distinct subtypes. Can Res. 2019;79(1):61–71.
78. Kwon S, Kim SS, Nebeck HE, Ahn EH. Immortalization of different breast epithelial cell types results in distinct mitochondrial mutagenesis. Int J Mol Sci. 2019;20(11):2813.
79. Ferone G, Lee MC, Sage J, Berns A. Cells of origin of lung cancers: lessons from mouse studies. Genes Dev. 2020;34(15–16):1017–32.
80. Kim HJ, Park JW, Lee JH. Genetic architectures and cell-of-origin in glioblastoma. Front Oncol. 2021. https://doi.org/10.3389/fonc.2020.615400.
81. Flowers BM, Xu H, Mulligan AS, Hanson KJ, Seoane JA, Vogel H, et al. Cell of origin influences pancreatic cancer subtype. Cancer Discov. 2021;11(3):660–77.
82. Geboes K, Hoorens A. The cell of origin for Barrett's esophagus. Science. 2021;373(6556):737–8.
83. Moeini A, Haber PK, Sia D. Cell of origin in biliary tract cancers and clinical implications. JHEPReport. 2021

Publisher's Note